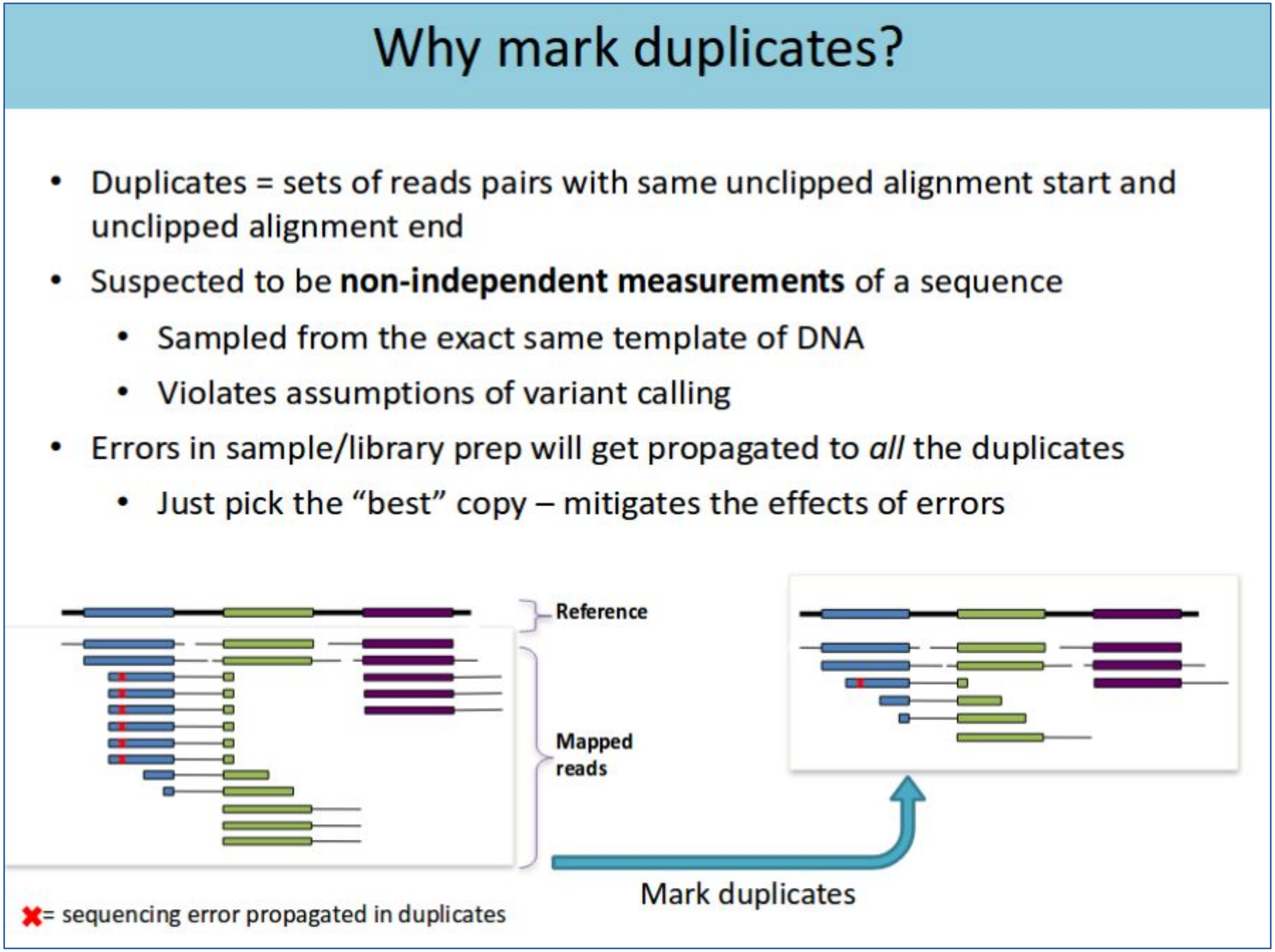
Module 1/5: Analyses ADN

- NGS Introduction
- Reads Quality Control
- Reads Cleaning
- Aligning reads on reference → *Hélène Touzet*
- Alignment parameters → *Hélène Touzet*
- Reads duplicates

→ **Practical #3**

Cleaning duplicated reads



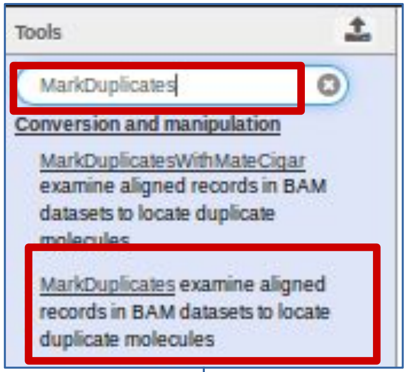
Why mark duplicates?

- Duplicates = sets of reads pairs with same unclipped alignment start and unclipped alignment end
- Suspected to be **non-independent measurements** of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- Errors in sample/library prep will get propagated to *all* the duplicates
  - Just pick the “best” copy – mitigates the effects of errors

Source : GATK Marking duplicates
https://software.broadinstitute.org/gatk/events/slides/1511/Presentations/GATKwh9-3-Marking_duplicates.pdf

Picard / MarkDuplicate

Additional information about Picard tools is available from Picard web site at *http://broadinstitute.github.io/picard/*



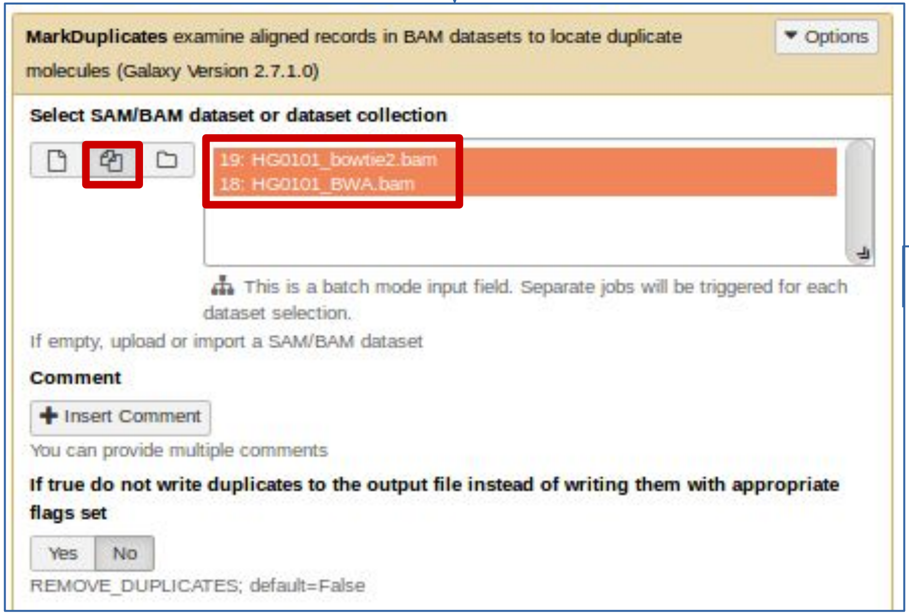
Tools

MarkDuplicates

Conversion and manipulation

MarkDuplicatesWithMateCigar examine aligned records in BAM datasets to locate duplicate molecules

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules



MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.7.1.0)

Select SAM/BAM dataset or dataset collection

19: HG0101_bowtie2.bam
18: HG0101_BWA.bam

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

If empty, upload or import a SAM/BAM dataset

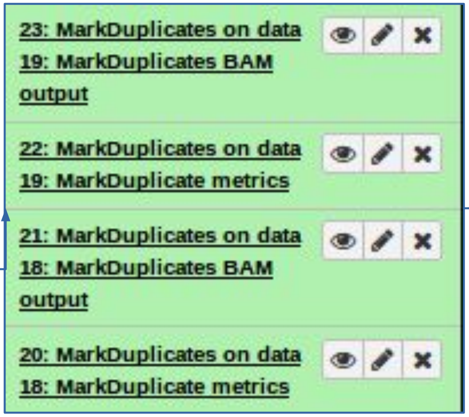
Comment

You can provide multiple comments

If true do not write duplicates to the output file instead of writing them with appropriate flags set

Yes [✓ No]

REMOVE_DUPLICATES; default=False

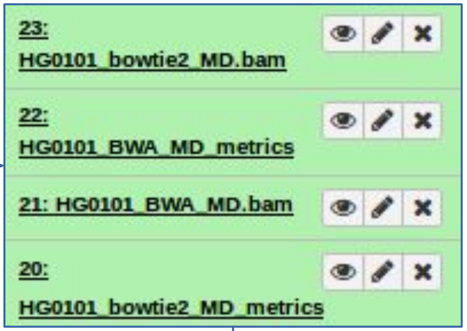


23: MarkDuplicates on data 19: MarkDuplicates BAM output

22: MarkDuplicates on data 19: MarkDuplicate metrics

21: MarkDuplicates on data 18: MarkDuplicates BAM output

20: MarkDuplicates on data 18: MarkDuplicate metrics

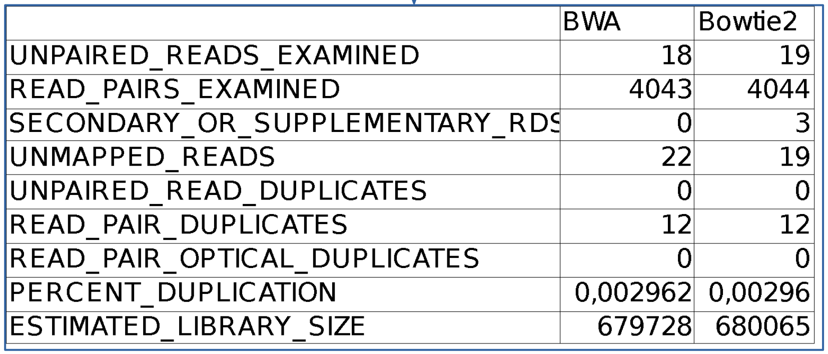


23: HG0101_bowtie2_MD.bam

22: HG0101_BWA_MD_metrics

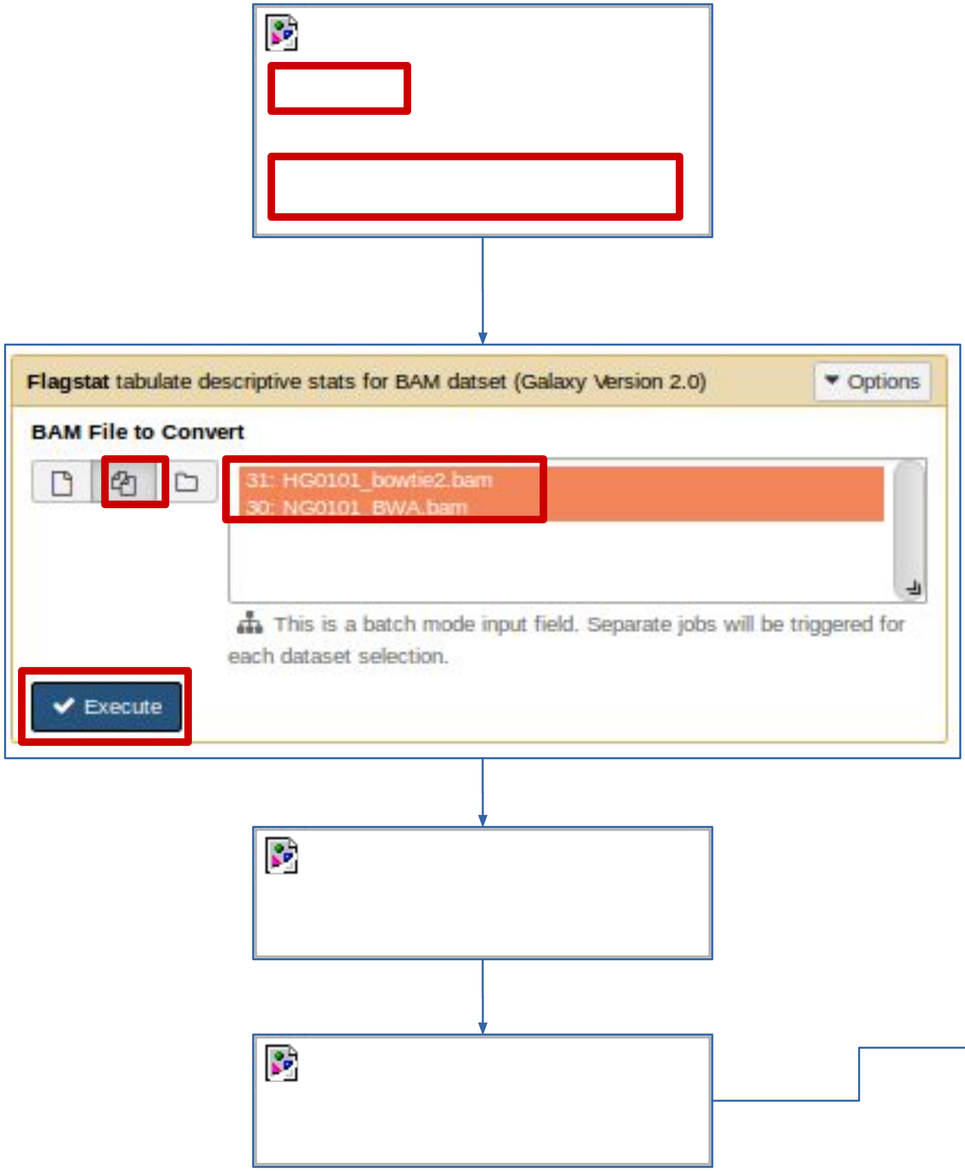
21: HG0101_BWA_MD.bam

20: HG0101_bowtie2_MD_metrics



	BWA	Bowtie2
UNPAIRED_READS_EXAMINED	18	19
READ_PAIRS_EXAMINED	4043	4044
SECONDARY_OR_SUPPLEMENTARY_RDS	0	3
UNMAPPED_READS	22	19
UNPAIRED_READ_DUPLICATES	0	0
READ_PAIR_DUPLICATES	12	12
READ_PAIR_OPTICAL_DUPLICATES	0	0
PERCENT_DUPLICATION	0,002962	0,00296
ESTIMATED_LIBRARY_SIZE	679728	680065

Alignment count : *samtools flagstat*



Flagstat tabulate descriptive stats for BAM datset (Galaxy Version 2.0)

BAM File to Convert

31: HG0101_bowtie2.bam
30: NG0101_BWA.bam

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

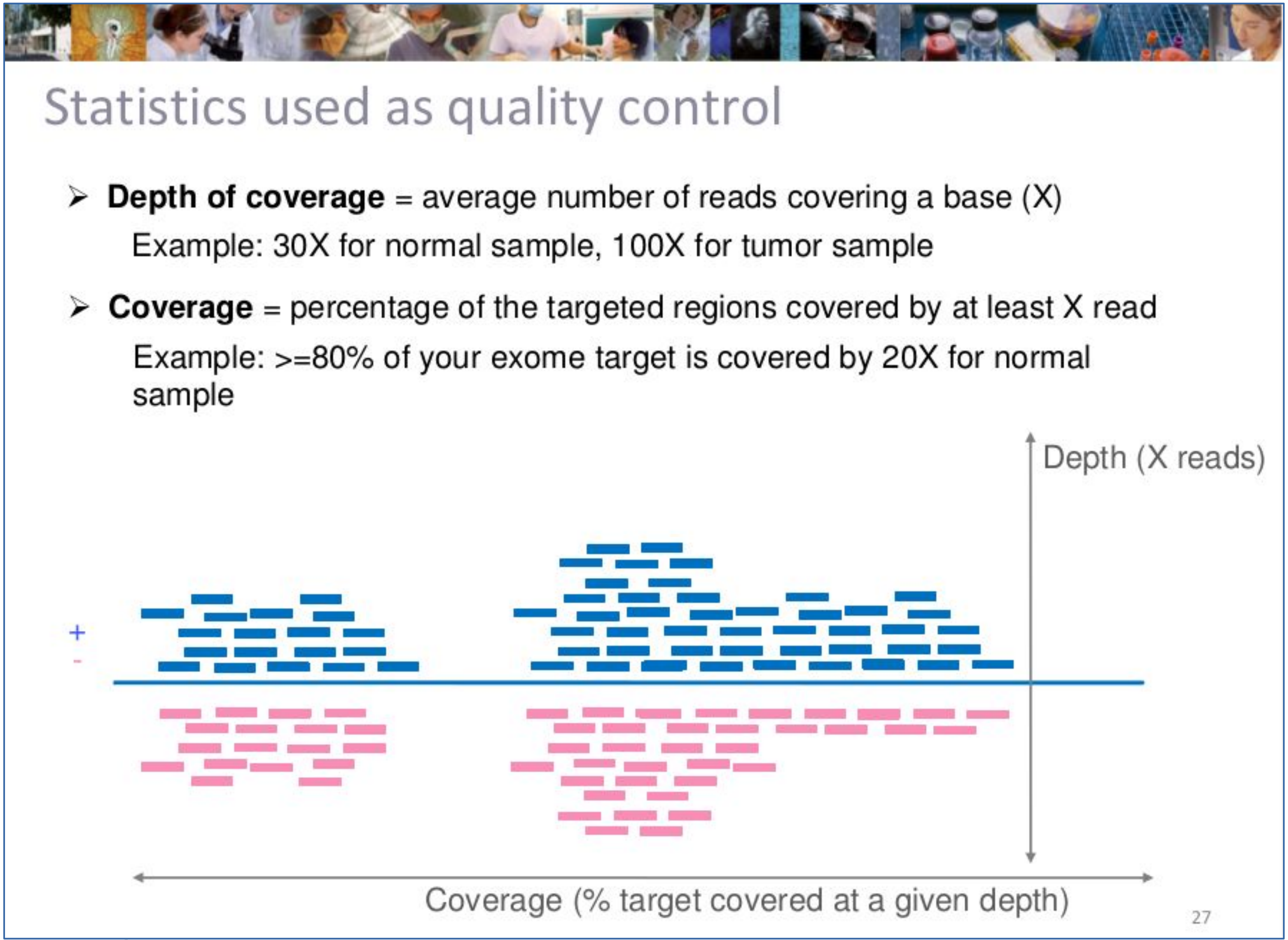
BWA

```
8129 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
3 + 0 supplementary
24 + 0 duplicates
8110 + 0 mapped (99.77%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
7980 + 0 properly paired (98.20%:-nan%)
8088 + 0 with itself and mate mapped
19 + 0 singletons (0.23%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Bowtie 2

```
8126 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
24 + 0 duplicates
8104 + 0 mapped (99.73%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
8074 + 0 properly paired (99.36%:-nan%)
8086 + 0 with itself and mate mapped
18 + 0 singletons (0.22%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Coverage and deepth of coverage



Statistics used as quality control

- **Depth of coverage** = average number of reads covering a base (X)

  Example: 30X for normal sample, 100X for tumor sample

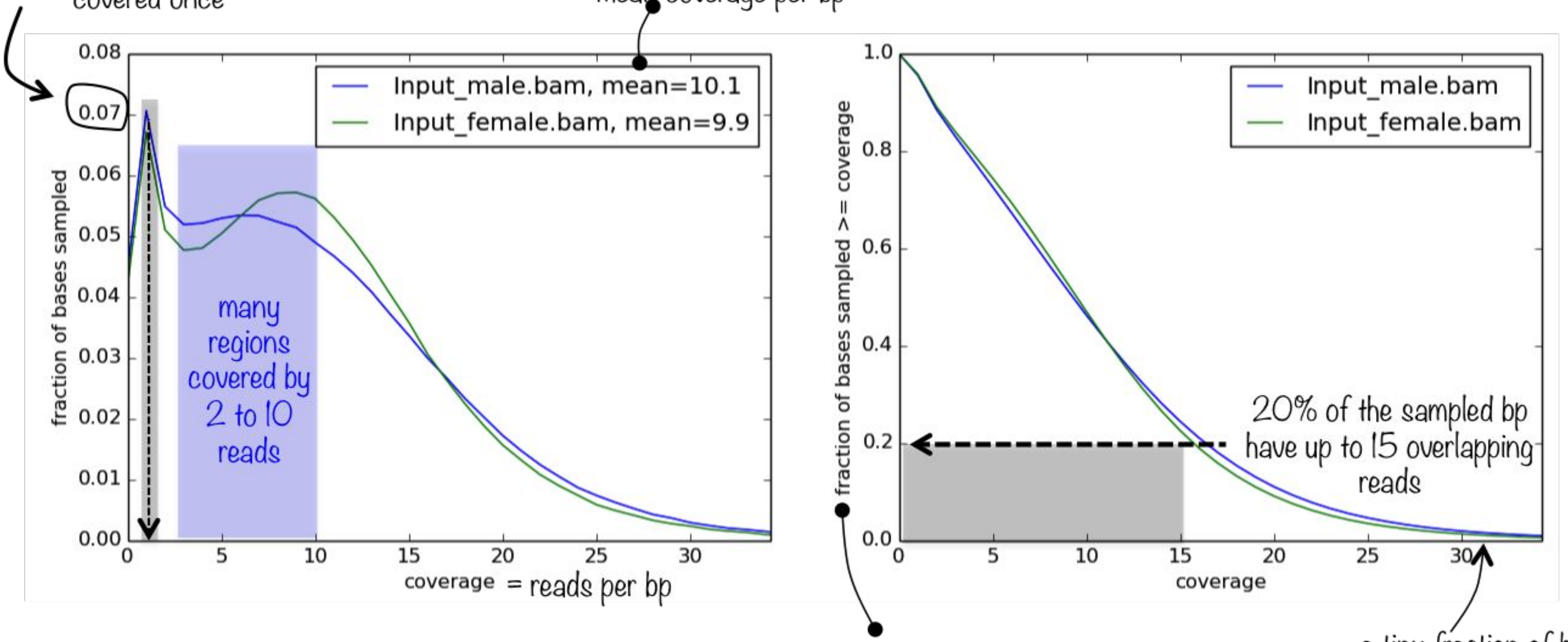
- **Coverage** = percentage of the targeted regions covered by at least X read

  Example: >=80% of your exome target is covered by 20X for normal sample

Source : Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016
http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf

Computing coverage and deepth of coverage
DeepTools2 / plotCoverage

Ramírez, Fidel and Ryan, Devon P and Grüning, Björn and Bhardwaj, Vivek and Kilpert, Fabian and Richter, Andreas S and Heyne, Steffen and Dündar, Friederike and Manke, Thomas (2016). deepTools2: a next generation web server for deep-sequencing data analysis. In Nucleic Acids Research, 44 (W1), pp. W160–W165



7% of the sampled bp are covered once

mean coverage per bp

many regions covered by 2 to 10 reads

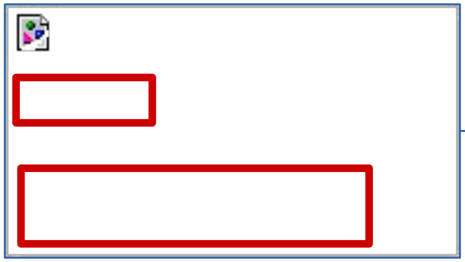
= reads per bp

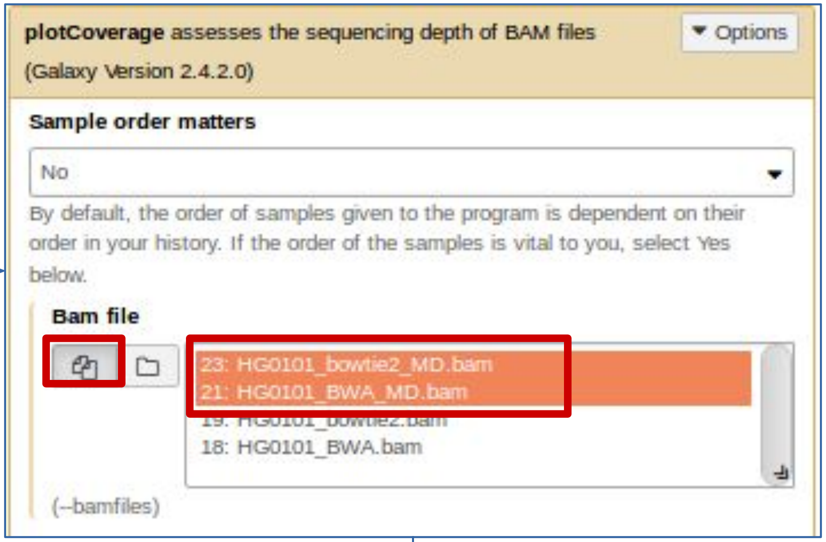
20% of the sampled bp have up to 15 overlapping reads

basically, a "reverse cumulative sum"

a tiny fraction of bp has more than 30 overlapping reads

DeepTools / Plot Coverage





plotCoverage assesses the sequencing depth of BAM files (Galaxy Version 2.4.2.0)

Sample order matters

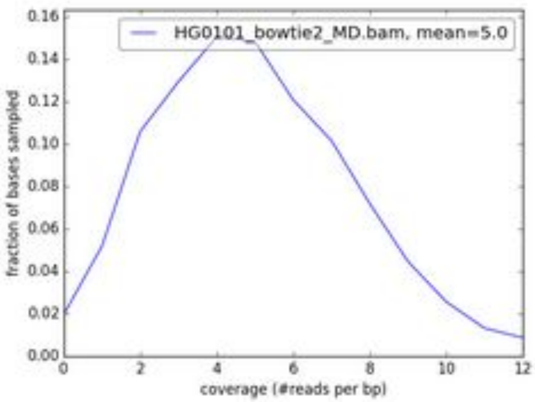
No ▾

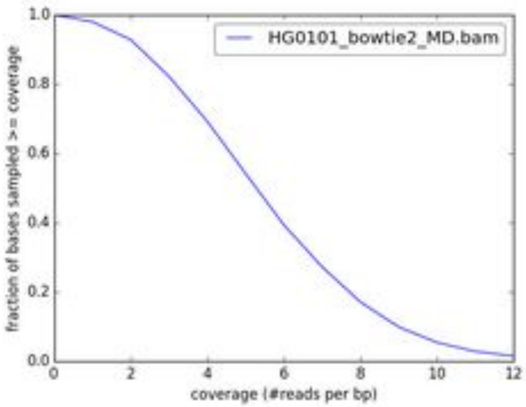
By default, the order of samples given to the program is dependent on their order in your history. If the order of the samples is vital to you, select Yes below.

Bam file

23: HG0101_bowtie2_MD.bam
21: HG0101_BWA_MD.bam
19: HG0101_bowtie2.bam
18: HG0101_BWA.bam

(--bamfiles)

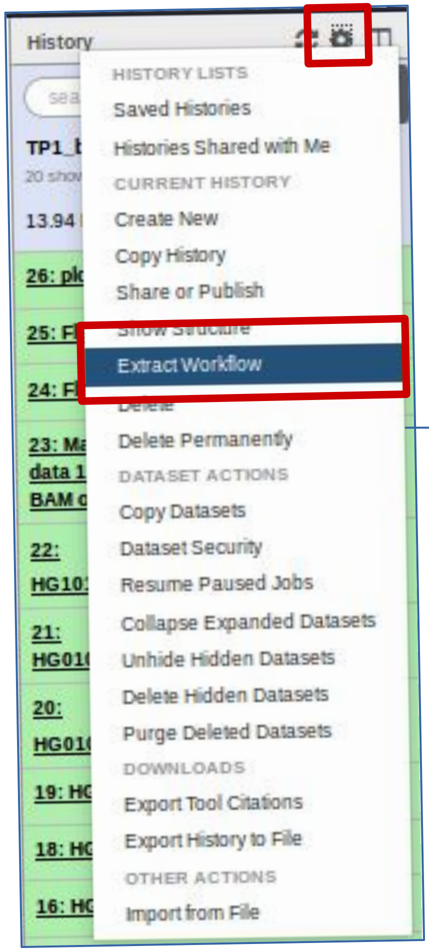


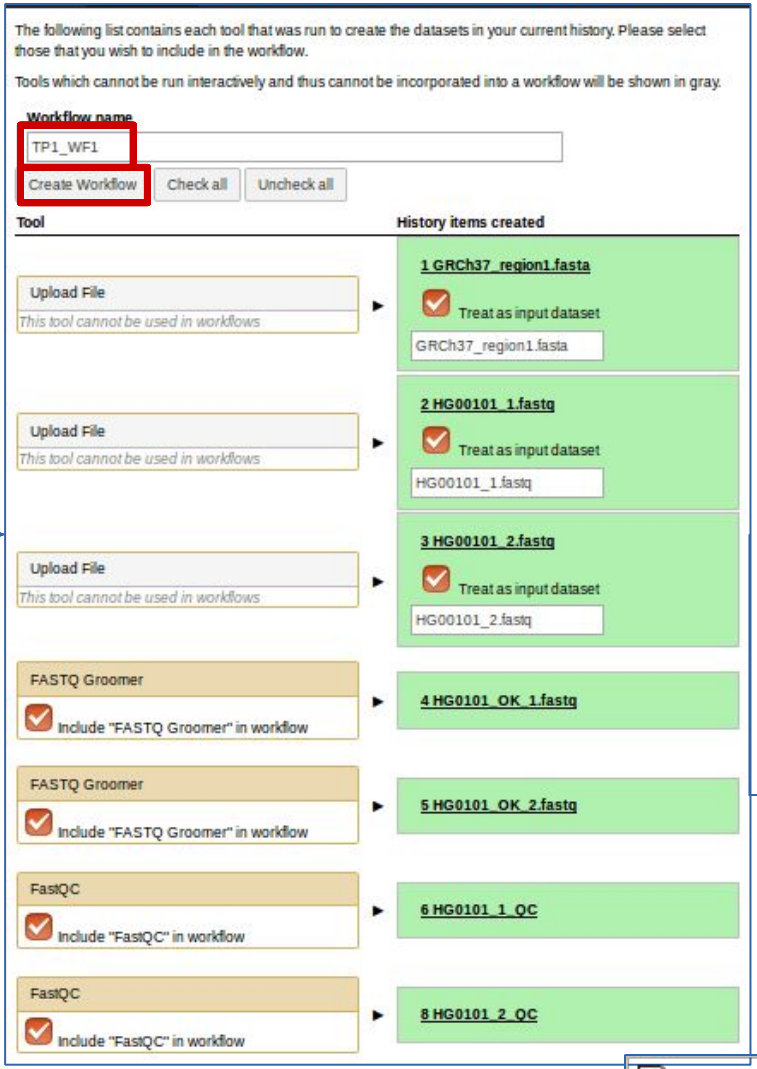


Galaxy *Workflow*

- Extract *workflow* from an history
- Modify *workflow*
- Execute *workflow* on new data
- Compare results from 2 *workflows* (in 2 histories)

Extract *Workflow* from the history of steps applied to the first sample





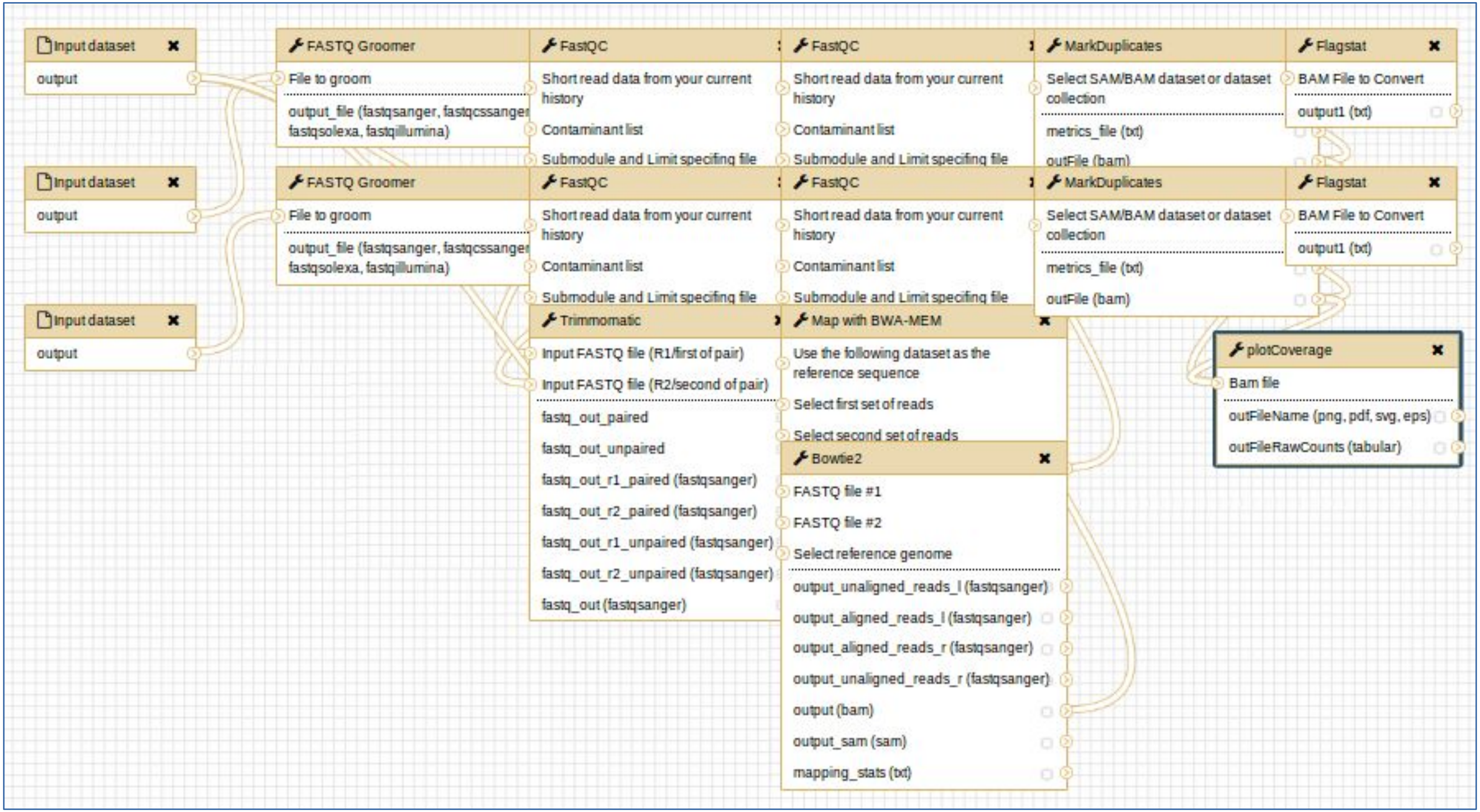
The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

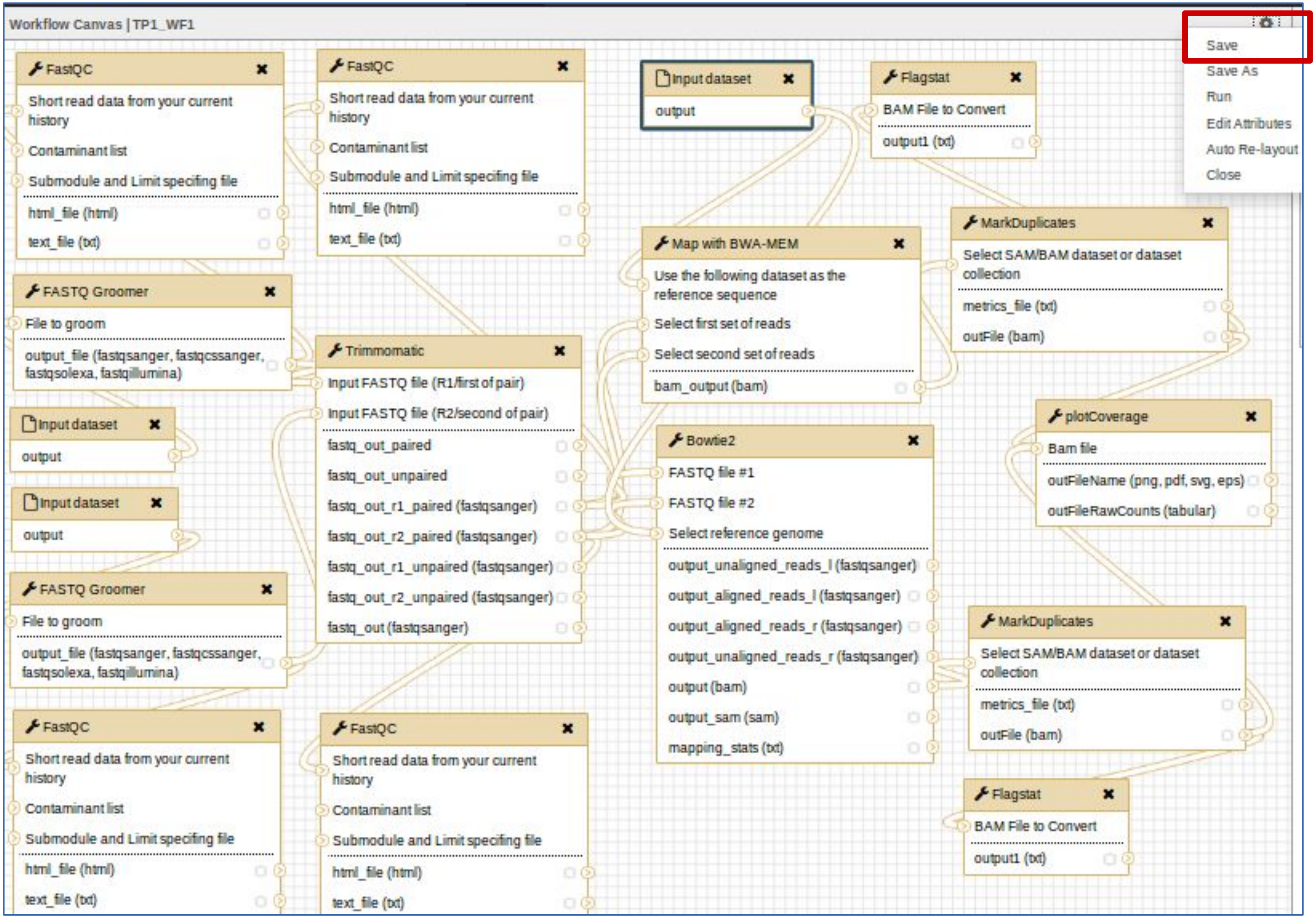
Workflow name: TP1_WF1

Tool	History items created
Upload File *This tool cannot be used in workflows*	1 GRCh37_region1.fasta ☑ Treat as input dataset GRCh37_region1.fasta
Upload File *This tool cannot be used in workflows*	2 HG00101_1.fastq ☑ Treat as input dataset HG00101_1.fastq
Upload File *This tool cannot be used in workflows*	3 HG00101_2.fastq ☑ Treat as input dataset HG00101_2.fastq
FASTQ Groomer ☑ Include "FASTQ Groomer" in workflow	4 HG0101_OK_1.fastq
FASTQ Groomer ☑ Include "FASTQ Groomer" in workflow	5 HG0101_OK_2.fastq
FastQC ☑ Include "FastQC" in workflow	6 HG0101_1_QC
FastQC ☑ Include "FastQC" in workflow	8 HG0101_2_QC

Visualize workflow

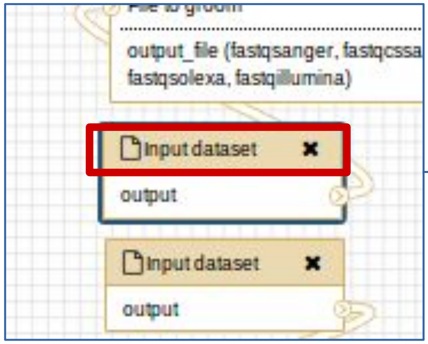

Input dataset
output
Input dataset
output
Input dataset
output
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
MarkDuplicates
Select SAM/BAM dataset or dataset collection
metrics_file (txt)
outFile (bam)
MarkDuplicates
Select SAM/BAM dataset or dataset collection
metrics_file (txt)
outFile (bam)
Flagstat
BAM File to Convert
output1 (txt)
Flagstat
BAM File to Convert
output1 (txt)
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
Map with BWA-MEM
Use the following dataset as the reference sequence
Select first set of reads
Select second set of reads
Bowtie2
FASTQ file #1
FASTQ file #2
Select reference genome
output_unaligned_reads_l (fastqsanger)
output_aligned_reads_l (fastqsanger)
output_aligned_reads_r (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
output_sam (sam)
mapping_stats (txt)
plotCoverage
Bam file
outFileName (png, pdf, svg, eps)
outFileRawCounts (tabular)

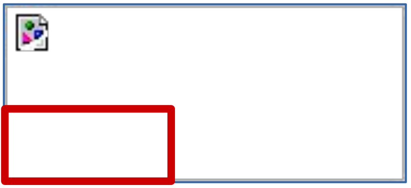
Modify workflow visualisation


Workflow Canvas | TP1_WF1
Save
Save As
Run
Edit Attributes
Auto Re-layout
Close
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Input dataset
output
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
Map with BWA-MEM
Use the following dataset as the reference sequence
Select first set of reads
Select second set of reads
bam_output (bam)
Flagstat
BAM File to Convert
output1 (txt)
MarkDuplicates
Select SAM/BAM dataset or dataset collection
metrics_file (txt)
outFile (bam)
Bowtie2
FASTQ file #1
FASTQ file #2
Select reference genome
output_unaligned_reads_l (fastqsanger)
output_aligned_reads_l (fastqsanger)
output_aligned_reads_r (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
output_sam (sam)
mapping_stats (txt)
plotCoverage
Bam file
outFileName (png, pdf, svg, eps)
outFileRawCounts (tabular)

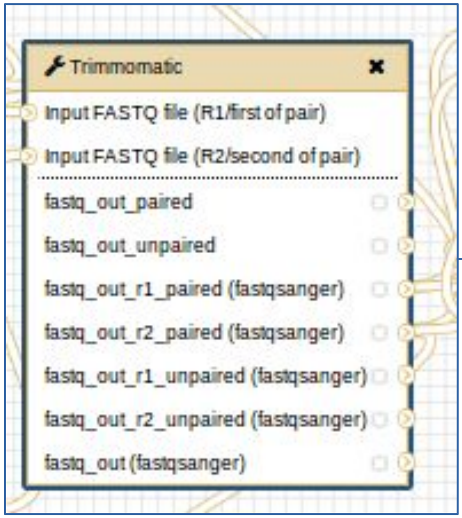
Modify some steps configuration

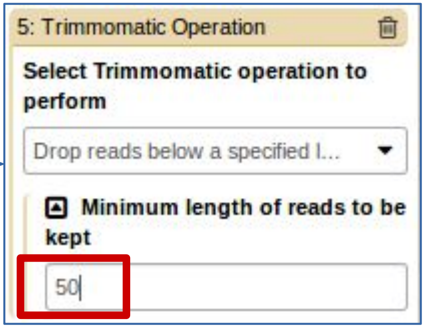
This WF uses 3 input files. Change box name to describe which data is required for each input : eg *Reference, Forward fastq, Reverse fastq*


File to groom
output_file (fastqsanger, fastqcssa
fastqsolexa, fastqillumina)
Input dataset
output
Input dataset
output



You can also change any parameter for example for *trimmomatic* step.


Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)



5: Trimmomatic Operation

Select Trimmomatic operation to perform

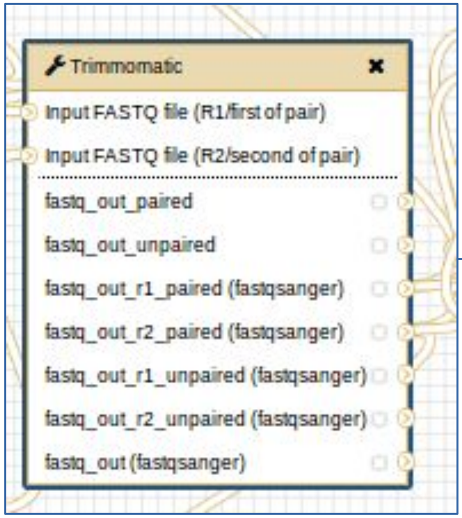
Drop reads below a specified l...

Minimum length of reads to be kept

50

Enable a parameter to be set at run time

Parameters for each tool will have the predefined values set in the workflow
You can modify this to enable any parameter to be set at run time.
Modify Trimmomatic so that Adapter are set at run time



Trimmomatic ✖

Input FASTQ file (R1/first of pair)

Input FASTQ file (R2/second of pair)

fastq_out_paired

fastq_out_unpaired

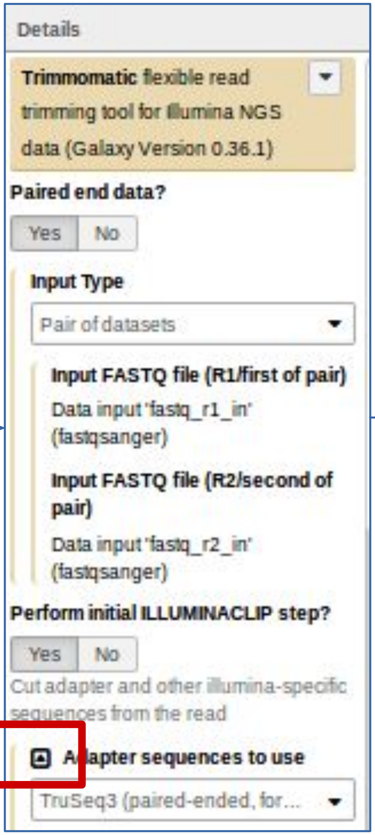
fastq_out_r1_paired (fastqsanger)

fastq_out_r2_paired (fastqsanger)

fastq_out_r1_unpaired (fastqsanger)

fastq_out_r2_unpaired (fastqsanger)

fastq_out (fastqsanger)



Details

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.1)

Paired end data?

Yes No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

Data input 'fastq_r1_in' (fastqsanger)

Input FASTQ file (R2/second of pair)

Data input 'fastq_r2_in' (fastqsanger)

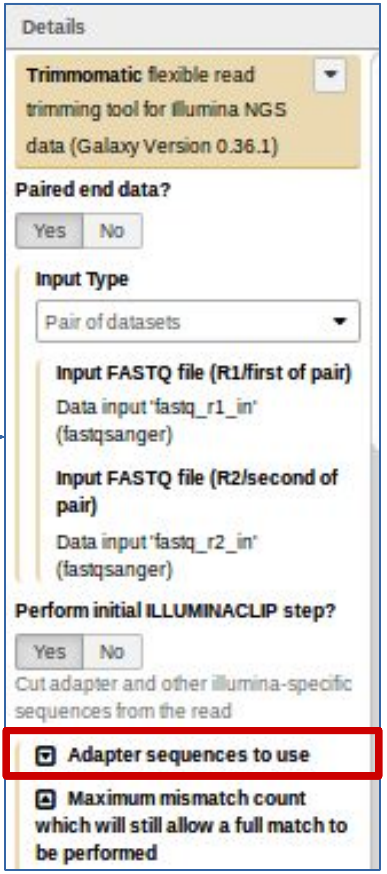
Perform initial ILLUMINACLIP step?

Yes No

Cut adapter and other illumina-specific sequences from the read

Adapter sequences to use

TruSeq3 (paired-ended, for...



Details

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.1)

Paired end data?

Yes No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

Data input 'fastq_r1_in' (fastqsanger)

Input FASTQ file (R2/second of pair)

Data input 'fastq_r2_in' (fastqsanger)

Perform initial ILLUMINACLIP step?

Yes No

Cut adapter and other illumina-specific sequences from the read

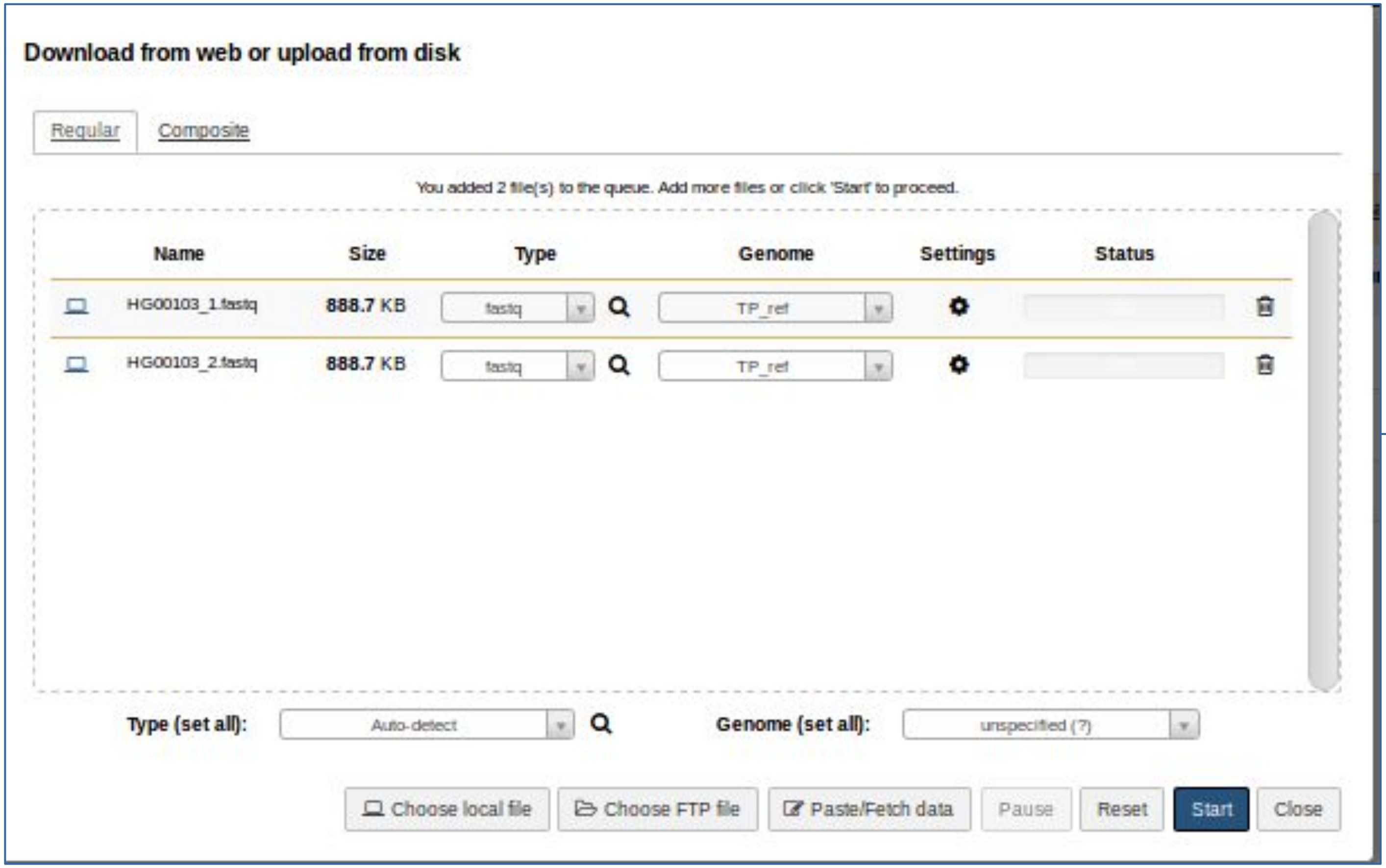
Adapter sequences to use

Maximum mismatch count which will still allow a full match to be performed

Do'nt forget to save your workflow !

Import new data for sample HG0103

Importe files HG0103_1.fastq and HG_0103_2.fastq



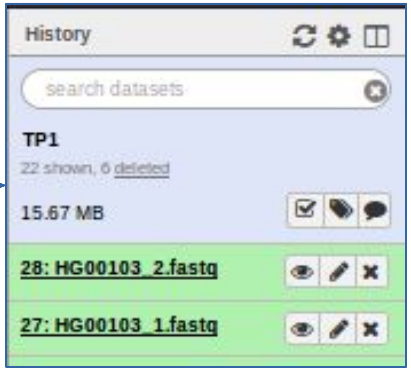
Download from web or upload from disk

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

	Name	Size	Type	Genome	Settings	Status
	HG00103_1.fastq	888.7 KB	fastq	TP_ref		
	HG00103_2.fastq	888.7 KB	fastq	TP_ref		

Type (set all): Auto-detect

Genome (set all): unspecified (?)



TP1

22 shown, 6 deleted

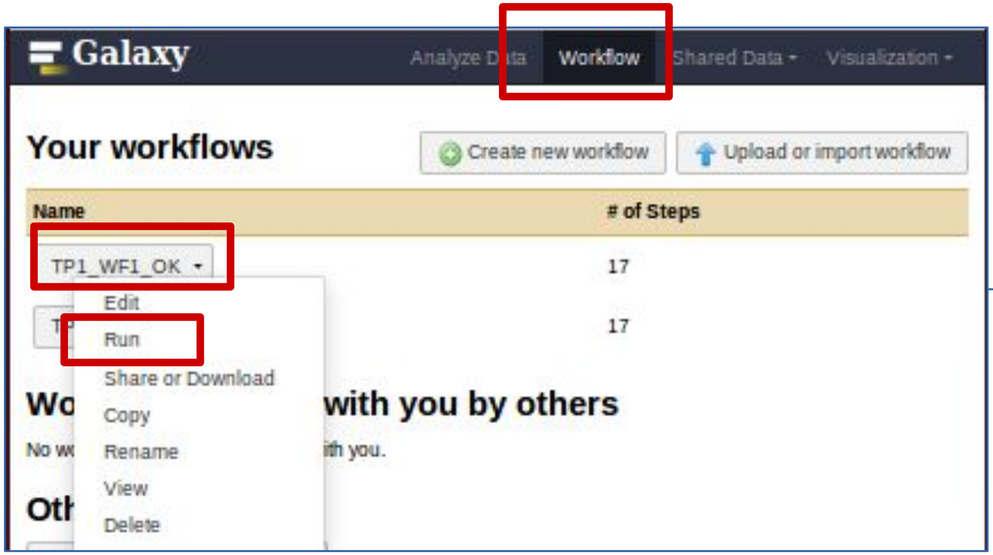
15.67 MB

28: HG00103_2.fastq

27: HG00103_1.fastq

Analyze these new data with the same workflow

Run the workflow with these new data



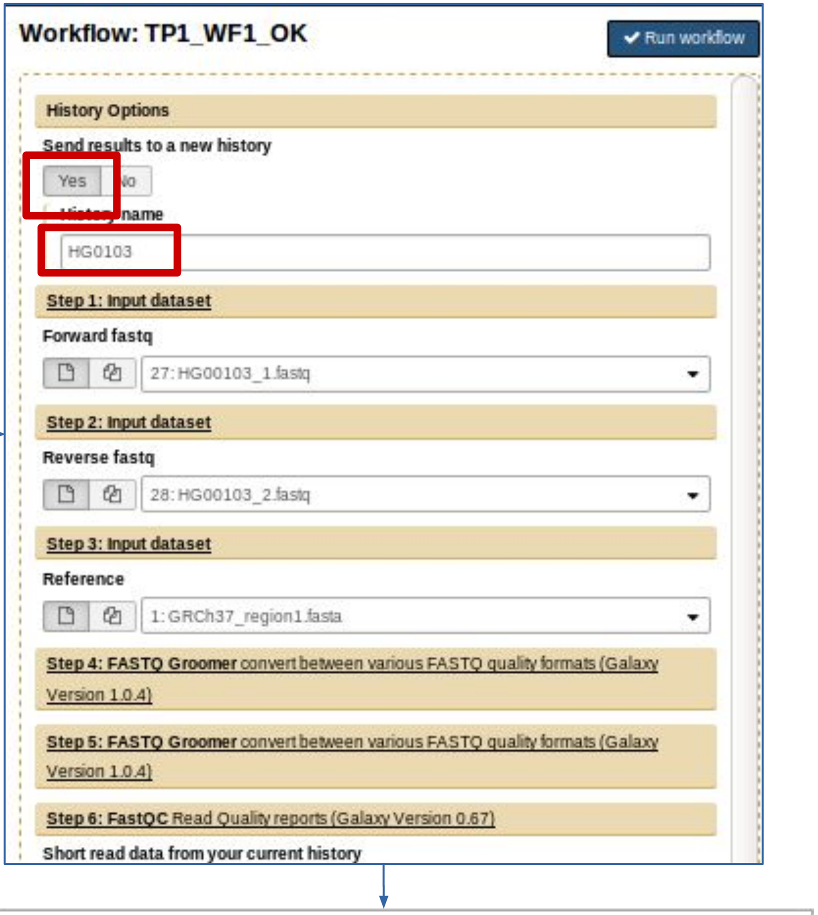
Your workflows

Name	# of Steps
TP1_WF1_OK ▾	17
T	17

Wo… with you by others

No w… …ith you.

Oth…



Workflow: TP1_WF1_OK

History Options

Send results to a new history: Yes

History name: HG0103

Step 1: Input dataset

Forward fastq: 27: HG00103_1.fastq

Step 2: Input dataset

Reverse fastq: 28: HG00103_2.fastq

Step 3: Input dataset

Reference: 1: GRCh37_region1.fasta

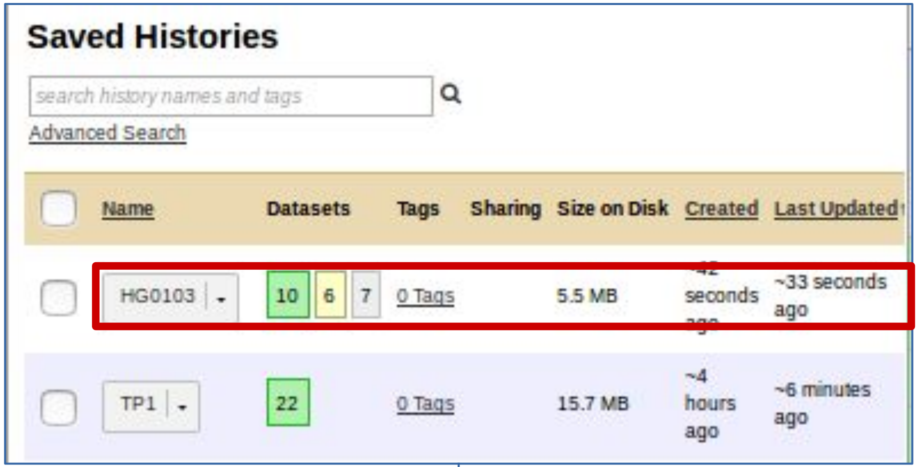
Step 4: FASTQ Groomer convert between various FASTQ quality formats (Galaxy Version 1.0.4)

Step 5: FASTQ Groomer convert between various FASTQ quality formats (Galaxy Version 1.0.4)

Step 6: FastQC Read Quality reports (Galaxy Version 0.67)

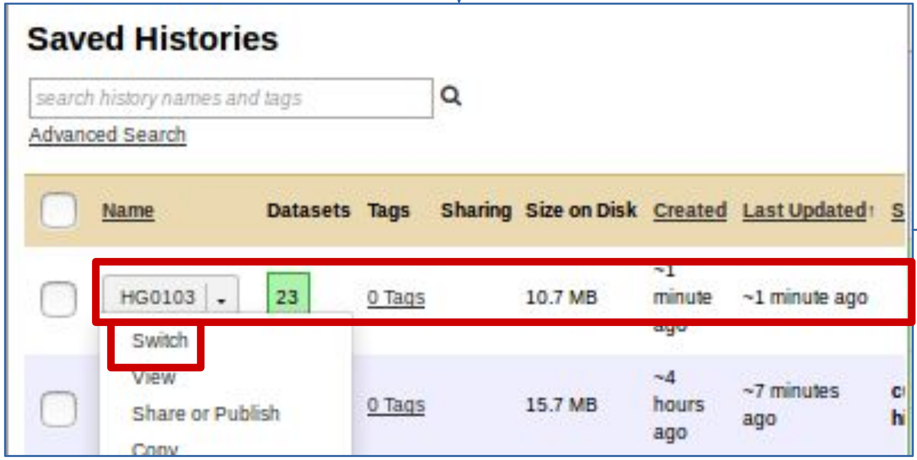
Short read data from your current history

Browse results



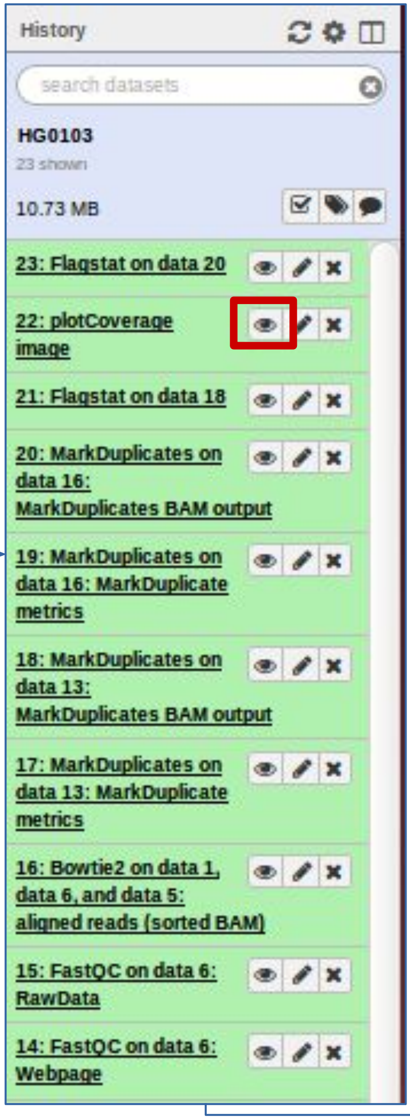
Saved Histories

□	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated
□	HG0103	10 6 7	0 Tags		5.5 MB	~42 seconds ago	~33 seconds ago
□	TP1	22	0 Tags		15.7 MB	~4 hours ago	~6 minutes ago



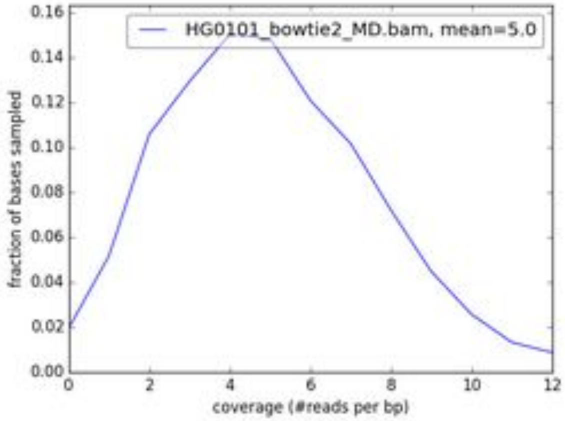
Saved Histories

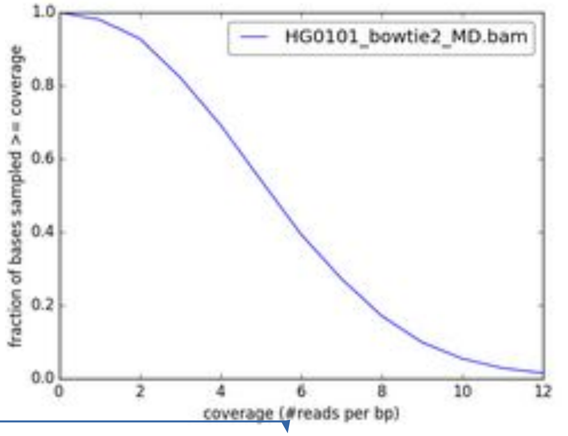
□	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	S
□	HG0103	23	0 Tags		10.7 MB	~1 minute ago	~1 minute ago	
□			0 Tags		15.7 MB	~4 hours ago	~7 minutes ago	c hi



HG0103
23 shown
10.73 MB

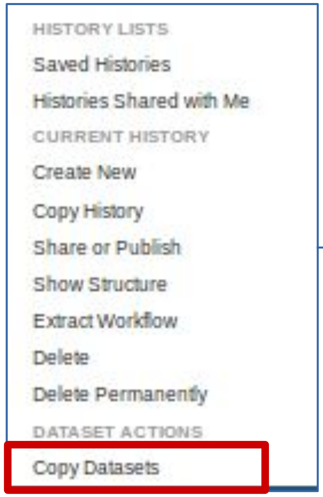
- 23: Flagstat on data 20
- 22: plotCoverage image
- 21: Flagstat on data 18
- 20: MarkDuplicates on data 16: MarkDuplicates BAM output
- 19: MarkDuplicates on data 16: MarkDuplicate metrics
- 18: MarkDuplicates on data 13: MarkDuplicates BAM output
- 17: MarkDuplicates on data 13: MarkDuplicate metrics
- 16: Bowtie2 on data 1, data 6, and data 5: aligned reads (sorted BAM)
- 15: FastQC on data 6: RawData
- 14: FastQC on data 6: Webpage

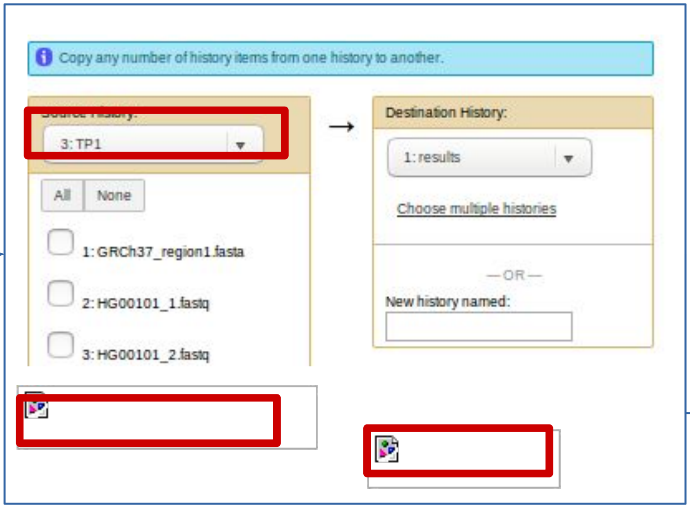




Gather BWA alignment results for the 2 samples 1/2

Create a new history named results
From history TP1 : Copy HG0101_BWA_MD.bam *dataset*




Copy any number of history items from one history to another.

Source History: 3: TP1

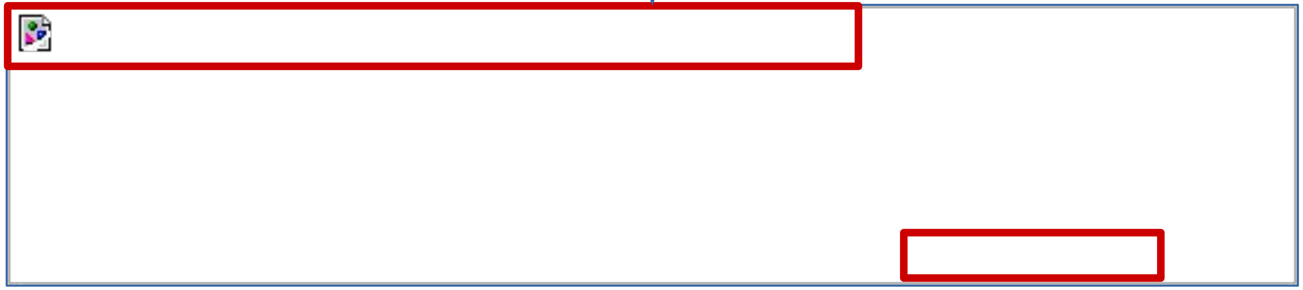
□ 1: GRCh37_region1.fasta
□ 2: HG00101_1.fastq
□ 3: HG00101_2.fastq

Destination History: 1: results

Choose multiple histories

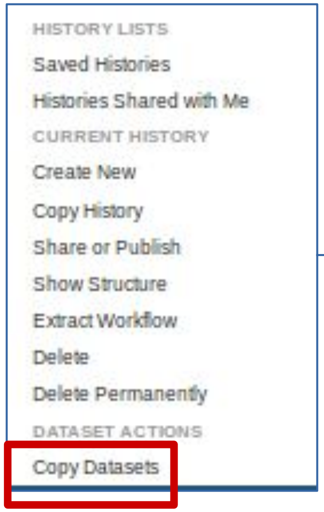
—OR—

New history named:

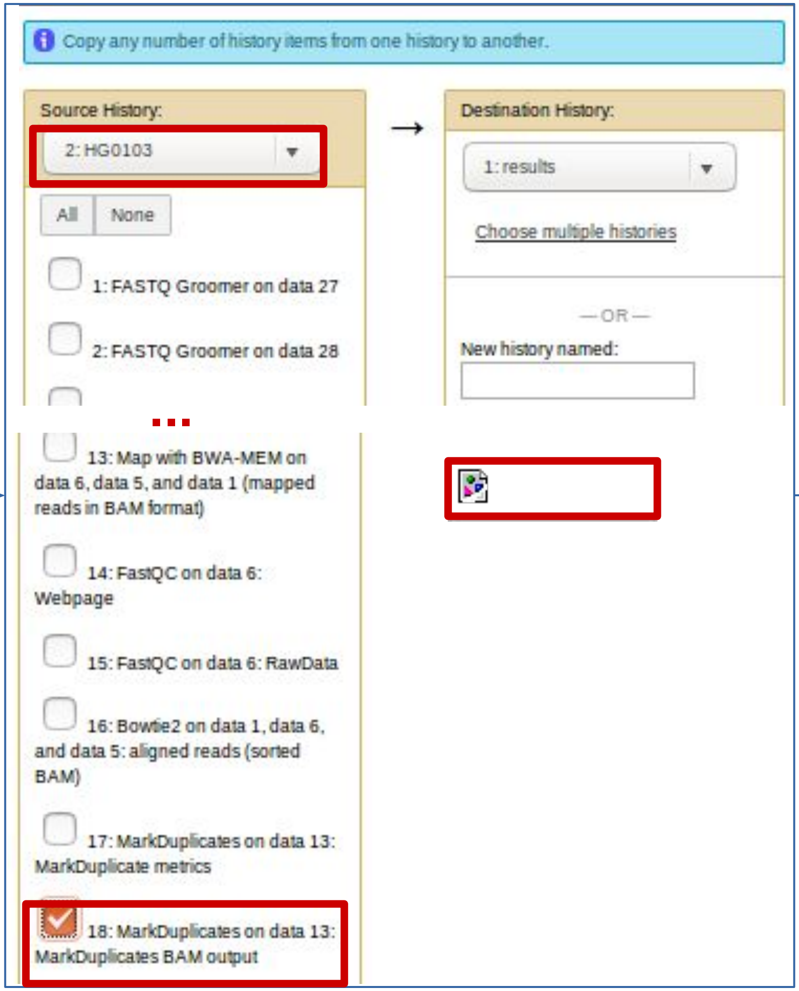


Gather BWA alignment results for the 2 samples 2/2

From history HG0103 : Copy *dataset*
« MarkDuplicates on data 13: MarkDuplicates BAM output »



HISTORY LISTS
Saved Histories
Histories Shared with Me
CURRENT HISTORY
Create New
Copy History
Share or Publish
Show Structure
Extract Workflow
Delete
Delete Permanently
DATASET ACTIONS
Copy Datasets



Copy any number of history items from one history to another.

Source History: 2: HG0103 → Destination History: 1: results

All None

☐ 1: FASTQ Groomer on data 27

☐ 2: FASTQ Groomer on data 28

...

☐ 13: Map with BWA-MEM on data 6, data 5, and data 1 (mapped reads in BAM format)

☐ 14: FastQC on data 6: Webpage

☐ 15: FastQC on data 6: RawData

☐ 16: Bowtie2 on data 1, data 6, and data 5: aligned reads (sorted BAM)

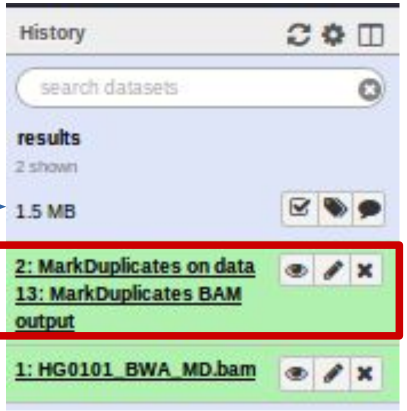
☐ 17: MarkDuplicates on data 13: MarkDuplicate metrics

☑ 18: MarkDuplicates on data 13: MarkDuplicates BAM output

Choose multiple histories

—OR—

New history named:



History

search datasets

results

2 shown

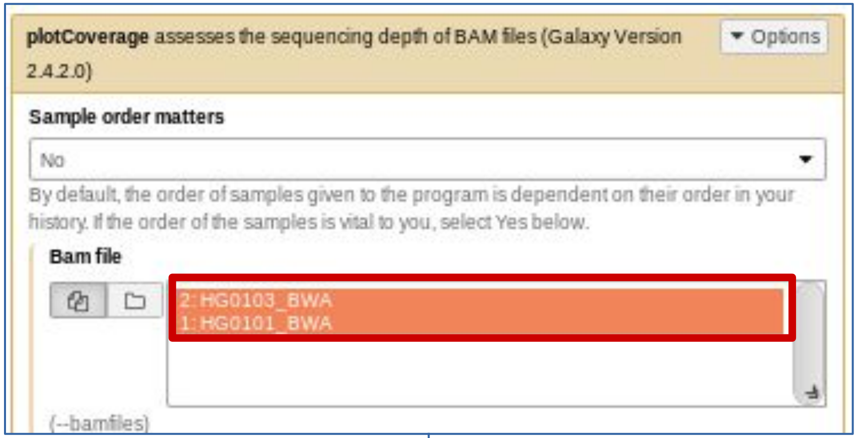
1.5 MB

2: MarkDuplicates on data 13: MarkDuplicates BAM output

1: HG0101_BWA_MD.bam

Visualize deepth of coverage for both samples

Rename *datasets*
Run *plotCoverage*



plotCoverage assesses the sequencing depth of BAM files (Galaxy Version 2.4.2.0)

Sample order matters

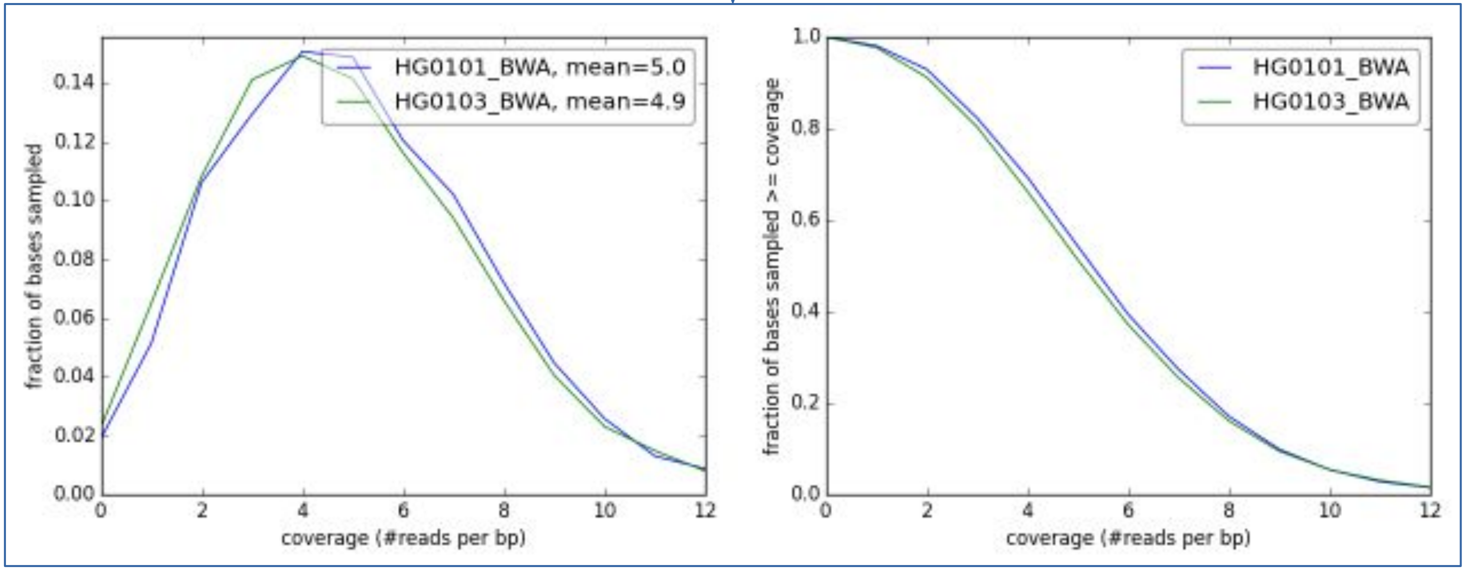
No

By default, the order of samples given to the program is dependent on their order in your history. If the order of the samples is vital to you, select Yes below.

Bam file

2: HG0103_BWA
1: HG0101_BWA

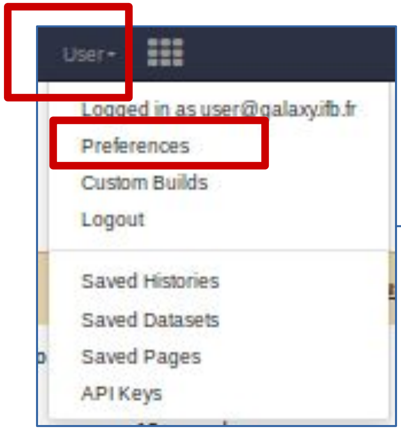
(--bamfiles)



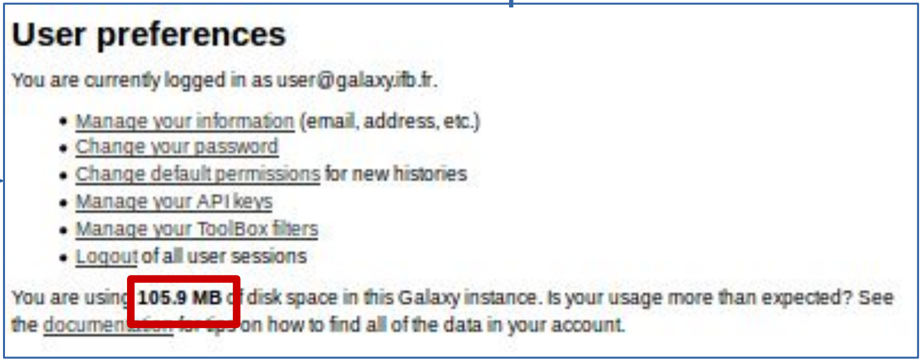
Galaxy – *Best Practices*

- Manage disk space
- Export analysis results (*datasets and histories*)
- Export / Import analysis protocoles (*workflow)*

Manage disk space



Global disk space



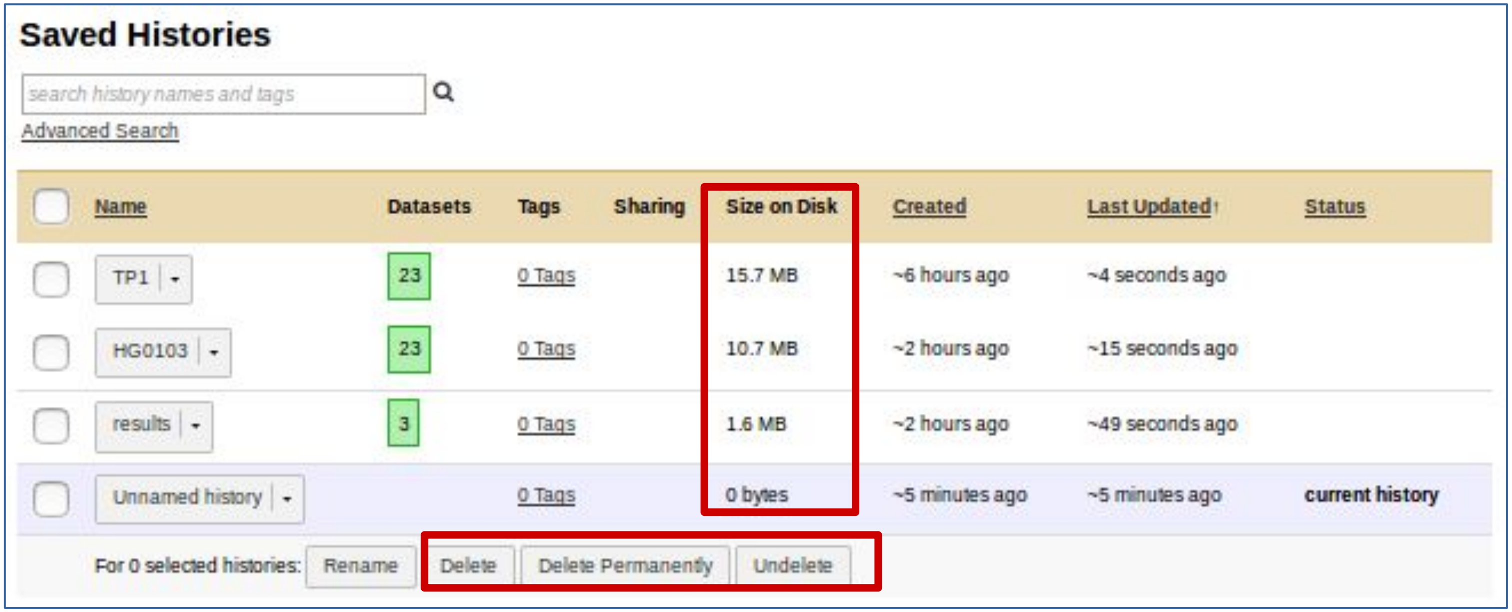
User preferences

You are currently logged in as user@galaxy.ifb.fr.

- Manage your information (email, address, etc.)
- Change your password
- Change default permissions for new histories
- Manage your API keys
- Manage your ToolBox filters
- Logout of all user sessions

You are using **105.9 MB** of disk space in this Galaxy instance. Is your usage more than expected? See the documentation for tips on how to find all of the data in your account.

Disk space per history



Saved Histories

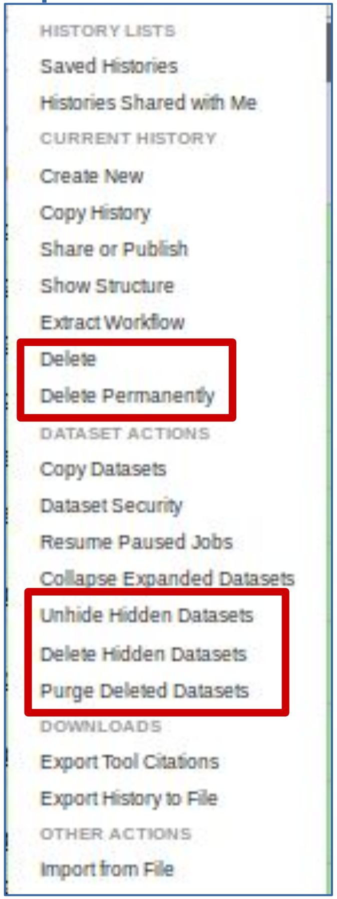
search history names and tags

Advanced Search

□	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
□	TP1 ▾	23	0 Tags		15.7 MB	~6 hours ago	~4 seconds ago	
□	HG0103 ▾	23	0 Tags		10.7 MB	~2 hours ago	~15 seconds ago	
□	results ▾	3	0 Tags		1.6 MB	~2 hours ago	~49 seconds ago	
□	Unnamed history ▾		0 Tags		0 bytes	~5 minutes ago	~5 minutes ago	current history

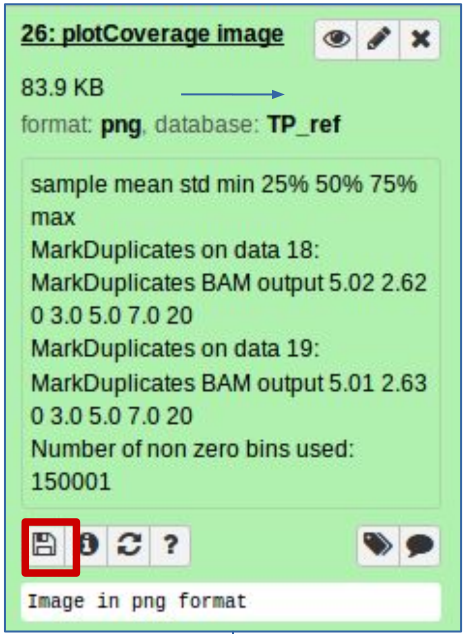
For 0 selected histories:

Disk space per *dataset*



Export analysis results : *datasets*

Image



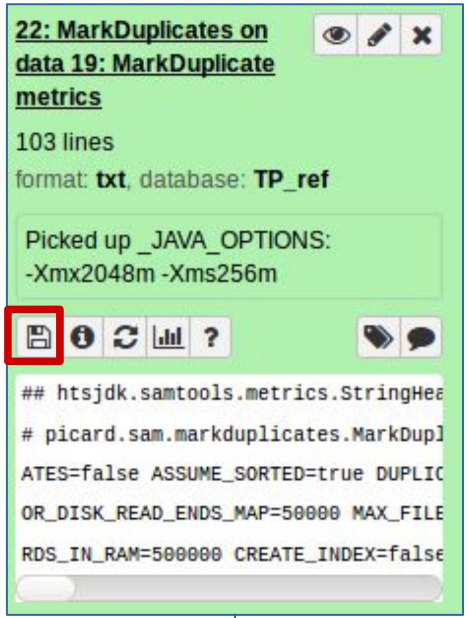
83.9 KB

format: **png**, database: **TP_ref**

sample mean std min 25% 50% 75% max
MarkDuplicates on data 18: MarkDuplicates BAM output 5.02 2.62 0 3.0 5.0 7.0 20
MarkDuplicates on data 19: MarkDuplicates BAM output 5.01 2.63 0 3.0 5.0 7.0 20
Number of non zero bins used: 150001

```
Image in png format
```

Text



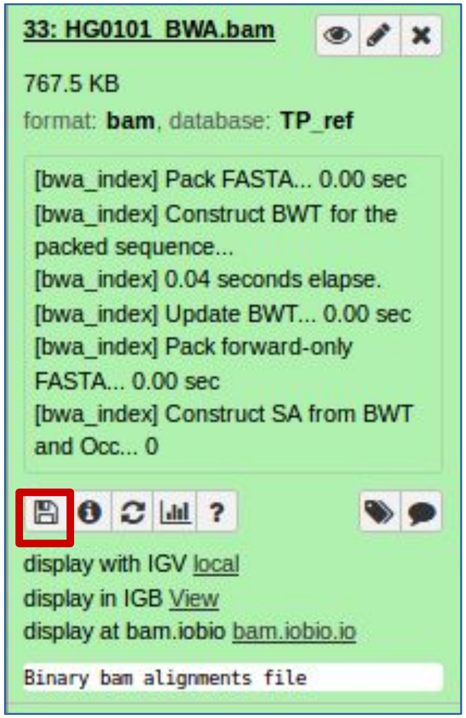
103 lines

format: **txt**, database: **TP_ref**

Picked up _JAVA_OPTIONS: -Xmx2048m -Xms256m

```
## htsjdk.samtools.metrics.StringHea
# picard.sam.markduplicates.MarkDupl
ATES=false ASSUME_SORTED=true DUPLIC
OR_DISK_READ_ENDS_MAP=50000 MAX_FILE
RDS_IN_RAM=500000 CREATE_INDEX=false
```

Bam



767.5 KB

format: **bam**, database: **TP_ref**

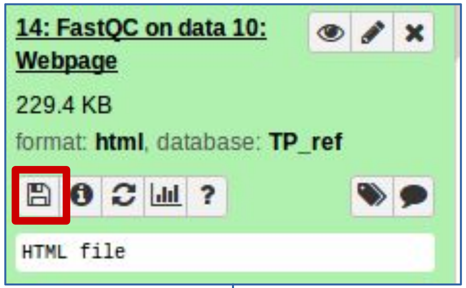
```
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.04 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forward-only FASTA... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0
```

display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io

```
Binary bam alignments file
```

Galaxy33-[HG0101_BWA.bam].bam

HTML + files

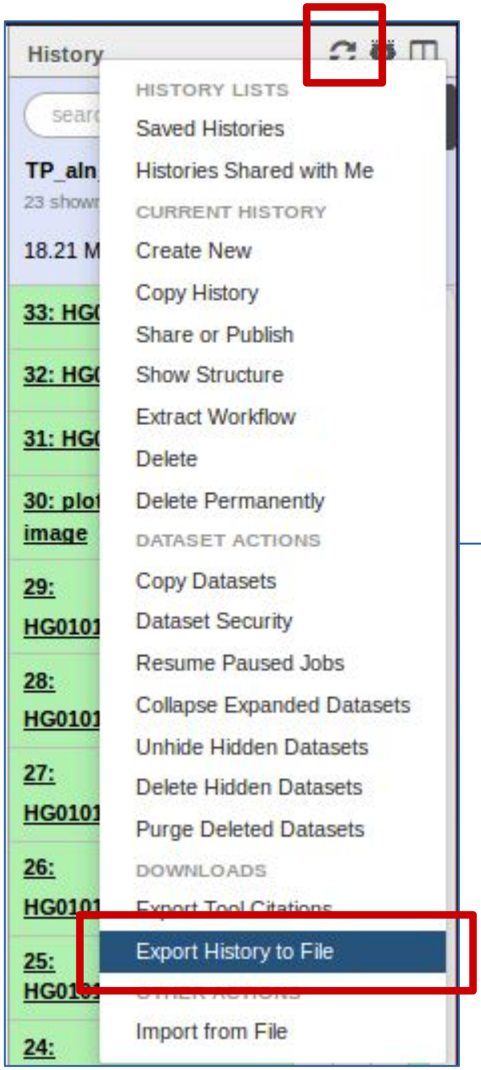


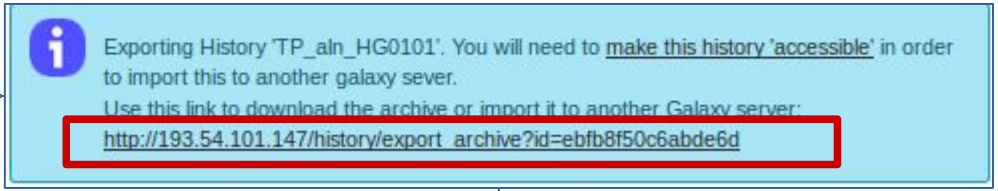
229.4 KB

format: **html**, database: **TP_ref**

```
HTML file
```

Export analysis results : *histories*

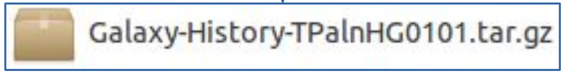




Exporting History 'TP_aln_HG0101'. You will need to make this history 'accessible' in order to import this to another galaxy sever.

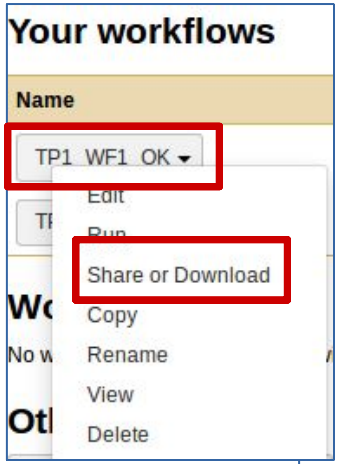
Use this link to download the archive or import it to another Galaxy server:

http://193.54.101.147/history/export_archive?id=ebfb8f50c6abde6d



Export / import analysis protocoles : *workflow*

Export


Your workflows

Import